3. MATERIALS AND METHODS

Computational biology is becoming one of the indispensable experimental tools in modern scientific research. The availability of high performing hardware and software technology facilitate better understanding of molecular insights at atomic level. For example, these computational technologies can extensively be used for protein, nucleotide sequence comparison, three dimensional structure prediction, and molecular interaction analysis.

3.1. Computational details

All computational work done in the present investigation has been carried out on high-performing server with 64 bit, Intel dual processor and 2 GB (Giga Bite) of RAM (Read Only Memory) configuration, and the operating system to handle software programs was Linux (Fedora core 2.4)

3.2. Computational screening of enzymes involved in the degradation of nitroaromatic compounds

The fundamental building blocks of life are proteins, which are the molecular machines responsible for virtually all of the chemical transformations that cells are capable of, in addition, much of the structure of a cell is made up of proteins. The part of the structure which is not made up of proteins is produced by enzymes (which are proteins). The sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments made using comparison tools with the data records available at public database.
3.3. Searching of databases for sequences

Entrez is an integrated text base search retrieval system used at NCBI for major databases including protein sequences, protein structures, taxonomy and others. The protein entries in Entrez search and retrieval system have been compiled from variety of sources including Swissprot. PDB translates from annotated coding regions in GenBank and reference sequences in addition to protein other related information is available via Entrez. The Entrez protein database is cross linked to the Entrez taxonomy database which allows us to find taxonomy information for the species from which protein sequence was derived.

GenBank is used to retrieve DNA sequences and protein sequences, in the form of FASTA format for further analysis. The data coding in GenBank is through never changing accession number, followed by a period and a version number. The version number starts at one, and increases by one each time the sequence changes. The second number is the GI number, where as the Swissprot databases codes by the nucleotide database unique identifier (Accession number; NID) or by PID. GenBank databases were accessed from www.Expasy.ch/sprot to retrieve amino acid sequences of the selected enzymes. The selected enzymes nitrobenzene nitroreductase, hydroxylaminobenzene mutase and 2-aminophenol 1, 6 dioxygenase sequences with accession numbers (gi 50236405, gi 2736268, gi 2444171) are collected in FASTA format.

3.3.1. Fast Sequence Alignment Program (FASTA)

FASTA is a fast sequence alignment program which is run from (http://www.ebi.ac.uk/fasta33/index.html ). FASTA stands for FAST-ALL, it can be used for fast protein sequence against all of the sequences in a database and return the most significant matches. FASTA scores only exact matches, and is also used for using online biological databank of NCBI, www.ncbi.nlm.nih.gov
Selected enzyme was searched which was at the left most comer of the NCBI home page. The query protein name or accession number was entered. Then the appropriate hit is selected from the list displayed, and from the FASTA the sequence is fetched and saved for further discussion. www.ebi.ac.uk/fasta33/index.html

3.3.2. Basic Local Alignment Search Tool

Basic Local Alignment Search Tool (BLAST) is widely used for searching protein and DNA databases for sequence similarities. The BLAST family of programs allows all combinations of DNA or protein query sequences with searches against DNA or protein databases. To achieve a marginally significant E-value of 0.05, a normalized score of-38 bits is necessary. The first reported sequence presents with 100% identity for a database sequence. The E value is very close to zero (for e.g. 3.9e-15, or 3.9 times ten to the power of minus 15, or 0.0000000000000015) suggesting that it is unlikely that this match is the result of chance (Altschul and Lipman, 1990).

**BlastP:** compares an amino acid query sequence against a protein sequence database.

**BlastN:** compares a nucleotide query sequence against a nucleotide sequence database.

**BlastX:** compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

**tblastn:** compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

**tblastx:** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The sequence is compared through the above programs to find out the most significant matches of our query sequence and results were analyzed. The resultant of BLASTP and BLASTN (www.ncbi.nlm.nih.gov/BLAST) and its most potential hits were analyzed and discussed.
3.4. Phylogenetic tree

Two types of trees are constructed taking branch lengths into consideration

a) Neighbor-joining method predicted by tree view online tool

b) UPGMA (unweighted pair group method with Arithmetic mean)

The tree view online tool is run by submitting the query protein sequence of enzymes to the phylogenetic analysis software, (http://tolweb.org/tree/phylogeny.html)

3.5. Primary Sequence analysis by PROTPARAM

PROTPARAM is a tool used to compute various physico-chemical properties for a protein sequence. The protein can either be specified as Swissprot/TrEMBL accession number or ID or in the form of raw sequence. The parameters computed by PROTPARAM include the molecular weight, theoretical pI, amino acid composition and atomic composition (http://www.expasy.org/tools/protparam.html).

3.6. Secondary structure prediction by GOR method

The GOR method was used to compute helices and sheets in the protein sequence. The protein was specified as Swissprot/TrEMBL accession number in the form of raw sequence. The secondary consensus sequences were predicted using http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html.

3.7. Homology modeling

The 3D structure Prediction of proteins can be done by Swiss Protein Data Bank Viewer (SPDBV) software. Deep view allows building models with specified conformations. This SPDBV software introduces tools for conformational analysis. The structural template was identified from the Protein Data Bank for the query sequence and loaded in Deep view. It can display more than one model at a time so the templates are superimposed to generate the structural alignments of the sequences. Swiss-PdbViewer (DeepView) has been developped since 1994 by Nicolas Guex. Swiss-PdbViewer is

36
tightly linked to SWISS-MODEL, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel.

Swiss-PdbViewer can also read electron density maps, and provides various tools to build into the density. In addition, various modeling tools are integrated and command files for popular energy minimization packages can be generated.

3.7.1. Prediction of 3D structure of selected enzyme by MODELLER

a. Preparing alignment files

The protein sequence was the enzyme, nitrobenzene nitroreductase, sequence fetched from PDB. It was copied and pasted in NCBI blast-p page by choosing PDB database in the BLASTP option. The structural sequence based on degree of similarity was selected and noted down its pdb id. The sequence which shows highest similarity with the selected enzyme sequence was copied and pasted in same note pad in which our target sequence was pasted and the file was saved. Then Clustal W software was chosen to upload the saved sequence. (www.ebi.ac.uk/clustalw/).

b. Preparing Atom files

The PDB id, which has got match with the enzyme sequence was taken and submitted to PDB database (access at www.rscb.org).

From the structural data opened the matched sequence structure was downloaded and was saved. (*.atm) file was prepared.

c. Preparing Script files

The standard format of script file was written which will be in note pad and the fields are filled required by the modeler. Then the file was saved giving the file name with the extension of *.top. All three files (*.ali, *.top and *.atm) are copied and specific path was given in DOS program like (C:\mod6v2\bin) command was given as "mod*.top"
(*indicates filename). The output file was created automatically by Modeller in bin directory approximately 30 min of analysis. The output file was copied and pasted in a folder and the same file is saved as *.pdb to visualize the enzyme 3D model created in Rasmol (visualization package)

3.7.2. Measuring model quality by PROCHECK analysis

The PROCHECK suite of programs provides a detailed check on the stereochemistry of a protein structure. PROCHECK is an online tool selected from www.Expasy.org or www.biochem.ucl.ac.uk/roman/procheck/procheck.html.

The predicted enzyme 3D model file was copied and pasted in Procheck tool which was runned to give output file. The output file shows the arrangement of amino acids regions in Ramachandran plot and the results were analyzed.

3.7.3 Measuring Model Quality by ERRAT Server

ERRAT is a protein structure verification algorithm that is especially well-suited for evaluating the progress of crystallographic model. This is extremely useful in making decisions about reliability. The model was submitted to ERRAT server www.doe.mbi.ucla.edu/people/errat in PDB format. The output file was observed as a graph.

3.8. Prediction of catalytic sites of the selected enzymes by E1DS Server

E1DS is a web server for enzyme catalytic site prediction. E1DS is designed for annotating enzyme sequences based on a repository of 1D signature. The selected enzyme sequence obtained from FASTA was copied and pasted in the sequence panel provided by the E1DS server. After 10 min, an output file was created which shows the catalytic sites or amino acids positions in the given sequence. E1DS is available at http://e1ds.ee.ncku.edu.tw/ and a mirror site can be found at http://e1ds.csbb.ntu.edu.tw/